Towards Mathematical Sciences

Vol. 4, No. 1, Spring and Summer 2024, pp. 19–31

DOI.10.22067/tmsj.2024.86497.1053

Expository Article

به سوی علوم ریاضی

https://tmsj.um.ac.ir

A COMPARATIVE STUDY ON THE RELATIONSHIP BETWEEN STATISTICS AND DATA SCIENCE

AMIR HOSSEIN GHATARI[1*], ELHAM TABRIZI[2], AND EHSAN BAHRAMI SAMANI[3]

[1]Department of Statistics, Faculty of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, IRAN
a.h.ghatari@aut.ac.ir
[2]Department of Mathematics, Faculty of Mathematical Science and Computer, Kharazmi University, Tehran, Iran
elham.tabrizi@khu.ac.ir
[3]Department of Statistics, Faculty of Mathematical Science, Shahid Beheshti University, Tehran, Iran
ehsan_ bahrami_ samani@yahoo.com

Abstract. The study of the relation between data science as an emergent phenomenon in 21st century and statistics, an original and applicable science, is one the most challenged topics by different researchers. Data science is dependent to some theoretical basics such as linear models, clustering and etc, which are from statistics. Some of the scientists believe that data science is begotten from statistics and consider it as statistics' offspring. Some others consider data science as a branch of statistics using modern computational techniques. Also, there is an uncommon idea that considers data science as a separated scientific field from statistics. In this paper, we recall theoretical basics of data science and bring up comments and interpretations of some well-known scientists about the fact that data science is not separate from statistics. Finally, we consider Statistics^{+} as another name of data science. Additionally, this study aims to explore the practical implications of this relationship by examining how advancements in data science continue to reshape and augment statistical methodologies. Through a comprehensive review of contemporary applications, we aim to illustrate the dynamic interplay between data science and statistics, illuminating their evolving synergy.

2010 Mathematics Subject Classification. 97K70, 97D10.
Keywords. Statistics, Statistics^{+} , Data Science, Machine Learning.
Date: Rceived 20-1-2024 Revised 29-2-2024 Accepted 28-5-2024 Available Online 27-9-2023

به سوی علوم ریاضی

دوره ۴، شماره ۱، بهار و تابستان ۱۴۰۳، صص. ۱۹–۳۱

DOI. 10.22067/tmsj.2024.86497.1053

مقاله علمی-ترویجی

https://tmsj.um.ac.ir

مطالعه‌ای تطبیقی بر ارتباط بین آمار و علم داده‌ها

امیرحسین قطاری*[1]، الهام تبریزی[2]، و احسان بهرامی سامانی[3]

[1] گروه آمار، دانشکده ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران.
a.h.ghatari@aut.ac.ir

[2] گروه ریاضی، دانشکده علوم ریاضی و کامپیوتر، دانشگاه خوارزمی، تهران، ایران.
elham.tabrizi@khu.ac.ir

[3]گروه آمار، دانشکده علوم ریاضی، دانشگاه شهید بهشتی، تهران، ایران.
ehsan_bahrami_samani@yahoo.com

چکیده. ارتباط میان علم داده‌ها به‌عنوان پدیده‌ای نوظهور در قرن ۲۱ با دانش ریشه‌دار و کاربردی آمار همواره مورد بحث و تبادل نظر میان پژوهشگران مختلف است. علم داده‌ها در میدان عمل، متکی به مبانی نظری مختلفی از جمله احتمالات، مدل‌های خطی و غیرخطی، رده‌بندی، روش‌های بازنمونه‌گیری، خوشه‌بندی و... است که جملگی از مبانی نظری علم آمار هستند. برخی از دانشمندان بر این باورند که علم داده‌ها برخاسته از آمار بوده و مولود این دانش سنتی است. برخی علم داده‌ها را شاخه‌ای از آمار می‌دانند که تکنیک‌های محاسباتی نرم افزاری نوین را به‌کار بسته است. همچنین، یک نظر نامرسوم نیز این است که علم داده‌ها فارغ از آمار بوده و بر آمار استوار نشده است. در این نوشتگان، ضمن بررسی پایه‌های نظری علم داده‌ها، بررسی شباهت مبانی نظری آن با آمار و نقد و بررسی نظرات برخی از صاحب‌نظران عصر حاضر این گزاره را که «علم داده‌ها عملا از آمار جدا نیست و همان آمار است با ابزارهای نوین» به‌عنوان خروجی این پژوهش مطرح کرده و لقب آمار+[۱] را برای علم داده‌ها درنظر می‌گیریم.

2010 Mathematics Subject Classification. 97K70, 97D10.

واژگان کلیدی. آمار، آمار+، علم داده‌ها، یادگیری ماشین.

تاریخ: دریافت ۱۴۰۲/۱۰/۳۰ بازنگری ۱۴۰۲/۱۲/۱۰ پذیرش ۱۴۰۳/۳/۸ انتشار برخط ۱۴۰۲/۷/۵

نحوه ارجاع به این مقاله: ا.ح. قطاری، ا. تبریزی، ا. بهرامی سامانی، مطالعه‌ای تطبیقی بر ارتباط بین آمار و علم داده‌ها، به سوی علوم ریاضی، ۴ (۱۴۰۳)، شماره ۱، ۱۹–۳۱.

۱. مقدمه

میان علم آمار و علم داده‌ها[1] ارتباط تنگاتنگی وجود دارد و بسیاری از روش‌های تحلیل داده از حوزه‌ی علم آمار اخذشده است؛ تا جایی که ارتباط علم داده‌ها و علم آمار موردبحث برخی پژوهشگران در قرن ۲۱ شده است. این‌که آیا علم داده‌ها شاخه‌ای از علم آمار است؟ یا این‌که بستگی به آمار ندارد و تنها از آمار استفاده می‌کند؟ همچنین، آیا علم داده‌ها بیشتر به علوم کامپیوتر نزدیک است تا آمار؟ ازجمله سؤالاتی هستند که در ضمن این مسئله مطرح‌شده‌اند. همواره نظرات مختلفی در مورد پاسخ به سؤالات فوق مطرح‌شده است. بهترین آغاز برای بررسی مسئله و پاسخ به سؤالات فوق و البته سؤالاتی از این قبیل، اشاره به تعاریف علم داده‌ها و آمار است که در نشریه انجمن آمار سلطنتی انگلستان و در مقاله دیگل[2] [۱۲] به‌کاررفته‌اند:

- آمار: این علم، دانش مطالعه‌ی مجموعه داده‌ها، تحلیل، تفسیر، نمایش و سازمان‌دهی آن‌هاست.
- علم داده‌ها: استخراج اطلاعات از داده‌ها است. فن‌ها و نظریه‌های مطرح‌شده در زمینه‌های مختلف علوم ریاضی، آمار و همچنین فناوری اطلاعات[3] را برای رسیدن به این هدف به کار می‌بندد.

به‌وضوح از تعریف این دو علم درمی‌یابیم که تقریباً هدف هر دو یک مقصد است با این تفاوت که آمار با ظاهری که امروزه مشاهده می‌کنیم در آغاز قرن بیستم و با مشارکت دانشمندانی چون فرانسیس گالتون[4]، رونالد فیشر[5]، کارل پیرسون[6] و.... شکلی نظام‌مند به خود گرفته است و در طی سالیان متمادی رشدی نظری و کاربردی داشته است؛ درحالی‌که علم داده‌ها با اهدافی مشابه آمار، در قرن ۲۱ پای به عرصه علوم نهاده است. به‌بیان‌دیگر، علم داده‌ها به‌طور مستقیم با فن‌آوری‌های نوین و پیشرفته ارتباط داشته و به استخراج مفاهیم از داده‌ها و تولید محصولات داده محور می‌پردازد.

برخی از دانشمندان، علم داده‌ها را زیرمجموعه‌ی آمار خوانده‌اند. ازجمله پژوهش‌های هم‌راستا با این نظر، می‌توان از گودمن[7] و همکاران [۱۷] و سینگپروالا[8] [۲۸] نام برد. همچنین کسلا[9] و همکاران [۷] معتقدند که آمار ریشه‌ی پیدایش دانشی است که به‌عنوان علم ساختاردهی، طبقه‌بندی و تحلیل داده شناخته می‌شود. در

[1]Data Science

[2]Diggle

[3]Information Technology

[4]Francis Galton

[5]Ronald Fisher

[6]Karl Pearson

[7]Goodman

[8]Singpurwalla

[9]Casella

یک گزارش اینترنتی برای نشریه‌ی پرایس اکونومیکس[10]، بهاردواج[11] [۲] به بررسی تفاوت‌های علم داده‌ها و آمار از دیدگاه‌های مختلف پرداخته است.

ازجمله موارد دیگری که از علل پیدایش علم داده‌ها محسوب می‌شود؛ نیاز مبرم به گسترش و تعمیم فن‌های محاسباتی آماری در تحلیل داده‌ها در دنیای امروز است. چمبرز[12] [۸] از قُدما در تحقیق و بررسی درباره‌ی نیاز به این مسئله بود. وی ازجمله پژوهشگرانی است که در قرن ۲۱ام در زمره‌ی موافقین این گزاره است که علم داده‌ها و مفاهیم مورداستفاده در آن ریشه‌ی آماری دارند. در پژوهشی به‌روزتر در مورد ارتباط میان علم داده‌ها و آمار، وایز[13] و ایکشتات[14] [۳۰]، به بررسی تأثیرات آمار در پیدایش علم داده‌ها پرداخته‌اند. آن‌ها معتقدند که آمار نقشی حیاتی در تشکیل علم داده‌ها و فراهم آوردن مفاهیم اولیه‌ی آن دارد. دیگر پژوهش توسط کارمیشل[15] و مارون[16] [۶] انجام‌شده است. آن‌ها این سؤال را که علم داده‌ها ریشه‌ی آماری دارد یا خیر را در ابعاد مختلف ازجمله نظرات سایر دانشمندان و حتی عملکرد در مسائل کاربردی، موردبررسی قراردادند. پژوهشگران متعددی درباره‌ی این‌که علم داده‌ها در مسیر به‌روزرسانی و تکامل آمار به وجود آمده است؛ تحقیق و بررسی کرده‌اند. برای اطلاعات بیشتر دراین‌باره به [۲۳، ۱۰، ۱۸، ۲۰، ۲۶] مراجعه شود.

در این مقاله به واکاوی جزئی‌تر اهداف علم داده‌ها، ابزار مورداستفاده در این علم و مقایسه‌ی آن با مفاهیم و ابزارهای آماری پرداخته و نظرات برخی از دانشمندان حال حاضر را درباره‌ی ارتباط این دو گرایش از علم موردنقد و بررسی قرار خواهیم داد. درنهایت می‌خواهیم به این سؤال پاسخ دهیم که آیا علم داده‌ها چیزی جدای از علم آمار است و یا تنها یک پدیده‌ی پیشرفته در ضمن آمار و متناسب با نیاز عصر حاضر است؟

هدف ثانویه‌ی این مقاله ارائه‌ی یک معرفی عمومی از آمار و علم داده‌ها به همراه برخی از مشاهیر و فعالان در این زمینه‌ها به نحوی است که فعالان علمی پژوهشی خارج از این دو حوزه را نیز همراه خود کرده و اطلاعات و منابع سودمندی درباره‌ی آمار و علم داده‌ها در اختیار این طیف از خوانندگان مقاله قرار دهد.

۲. علم داده‌ها، ظهور، اهداف و ابزارها

۱.۲. ظهور علم داده‌ها. عبارت علم داده‌ها بیش از دو دهه است که موجودیت دارد. با استناد به گزارش منتشرشده توسط جیل پرس[17] [۲۵]، ویلیام کلیولند[18] اولین کسی است که اصطلاح علم داده‌ها را در سال

[10] priceonomics

[11] Bhardwaj

[12] Chambers

[13] Weihs

[14] Ickstadt

[15] Carmichael

[16] Marron

[17] Press

[18] William Cleveland

۲۰۰۱ مطرح کرده است. وی در مقاله [۹] «علم داده‌ها: برنامه‌ای برای گسترش جنبه‌های فنی در رشته آمار» پیشنهاد کرد که علم داده‌ها به‌عنوان رشته‌ای مستقل شناخته شود . کلیولند این رشته‌ی جدید را مرتبط با علوم کامپیوتر و داده‌کاوی می‌دانست. وی بر این باور بود که منافع استفاده از یک تحلیلگر داده محدود است. چون مهندسین کامپیوتر شناخت کمی از روش‌های کار با داده دارند و دانش محاسباتی متخصصین آمار هم محدود است؛ بنابراین تلفیق این دو گروه می‌تواند منجر به نوآوری‌های زیادی شود. گروه‌های علم داده‌ها باید استادانی داشته باشد که بتوانند دانش داده‌ها را با دانش محاسبات تلفیق کنند.

۲.۲. اهداف علم داده‌ها. پیشرفت فنّاوری در روش‌های ذخیره‌سازی داده‌ها در عصر حاضر، منجر به پیدایش پایگاه‌های داده‌های بزرگ [۱۹] شده است. از طرف دیگر، پردازش و تحلیل داده‌های حجیم[۲۰] نظیر داده‌های هواشناسی، زیستی یا مالی به دلیل حجم و معمولاً بعد بالا، امر ساده‌ای نیست.

تحلیل داده‌های حجیم یک نیاز حیاتی در دنیای تجارت و همچنین علوم اجتماعی است [۲۴]. درواقع از نیازهای اساسی که می‌توان آن را یکی از علل پیدایش علم داده‌ها عنوان کرد؛ همین مسئله‌ی تحلیل داده‌های حجیم است. بر اساس نظر کلیولند، آمار سنگ بنای نظری برای علم داده‌ها را فراهم می‌کند و فنّاوری نیز به کمک آمار آمده و تحلیل‌های دقیق‌تر و متناسب با نیاز روز را فراهم می‌آورد. بر اساس همین نظر نیز می‌توان گفت که علم داده‌ها درواقع آمار است که بازوان اجرایی آن فن‌های مدرن در دنیای مهندسی کامپیوتر است.

۳.۲. ابزارهای علم داده‌ها. در این زیربخش به مروری بر ابزارها و فن‌های مرسوم در علم داده‌ها پرداخته می‌شود. از یک‌سو، مشخص می‌شود که علم داده‌ها خاستگاه نظری خود را از چه منبعی تهیه‌کرده و از طرف دیگر نشان می‌دهد که علم داده‌ها عملاً تلفیق چند رشته‌ی علمی نیست و درواقع آمار است که با استفاده از ابزارهای محاسباتی پیشرفته در عصر حاضر دست به ارائه‌ی نتایج می‌زند. این ابزارها عبارت‌اند از:

- رگرسیون خطی[۲۱]، مبحثی که گروه تحقیقاتی گالتون نخستین بار در اواخر قرن ۱۹ میلادی بدان پرداخته‌اند.

- مدل‌های خطی تعمیم‌یافته[۲۲]، این مقوله در ذیل مبحث کلاس‌بندی داده‌ها بوده که نخستین بار تحلیل این نوع از داده‌ها توسط کارل پیرسون در آغاز قرن ۲۰ میلادی مطرح گشت (اگرستی[۲۳] [۱] فصل ۱۶، بخش اول).

---

[19] Big Data

[20] Huge Data

[21] Linear Regression

[22] Generalized Linear Models

[23] Agresti

- درخت تصمیم[24]، این مبحث نیز یک شاخه‌ی نظری در علم آمار است که در مسائل رده‌بندی و رگرسیونی می‌توان از آن استفاده کرد. برای اطلاعات بیشتر در مورد درخت تصمیم به بریمن [25] و همکاران [۴] و ون وینترفلت [26] و ادواردز [27] [۲۹] مراجعه شود.

- در نوعی پیشرفته‌تر از درخت تصمیم، جنگل تصادفی[28]، کیسه‌بندی[29] و روش تقویتی[30]، سه روش مبتنی بر یادگیری گروهی[31] هستند که برای ساختن مدل‌های پیشگویی‌کننده از درخت‌های تصمیم استفاده می‌کنند. در اکثر روش‌های سنتی آمار فرض می‌شود که تعداد متغیرهای توضیحی کمتر از تعداد نمونه‌های مستقل است. بنابراین، ظهور مجموعه داده‌های با بعد بالا[32] که در آن‌ها تعداد متغیرهای توضیحی بیشتر از تعداد نمونه‌های مستقل است، به‌طور قابل‌توجهی جریان اصلی بسیاری از تحقیقات آماری را به خود اختصاص داده است. جنگل‌های تصادفی یکی از جدیدترین روش‌های یادگیری با ناظر[33] است که محققین، مطالعات زیادی را پیرامون این روش برای بهبود عملکرد پیش‌بینی متغیر پاسخ، هنگام مواجهه با داده‌های با بعد بالا، انجام داده‌اند. مجموعه اطلاعات در مورد این نوع از یادگیری آماری در هستی[34] و همکاران [۱۹] در دسترس است.
- ماشین بردار پشتیبان[35]، این نیز یک مبحث در کلاس‌بندی داده‌هاست که عملاً مباحثی نظری در آمار دارد. مجموعه اطلاعات در مورد این نوع از کلاس‌بندی، منشأ و معرف آن در هستی[36] و همکاران [۱۹] در دسترس است.

---

[24] Decision Tree

[25] Breiman

[26] Von Winterfeldt

[27] Edwards

[28] Random Forest

[29] Bagging

[30] Boosting

[31] Ensemble Learning

[32] High Dimensional Data

[33] Supervised Learning

[34] Hastie

[35] Support Vector Machine

[36] Hastie

- خوشه‌بندی[۳۷]، نخستین بار ارائه‌شده توسط دریور[۳۸] و کروبر[۳۹] [۱۴]، بحثی مجزا با گونه‌های بسیار متنوع در علم آمار است. تفاوت آن با کلاس‌بندی آمار این است که در کلاس‌بندی برچسب طبقات داده‌ها مشخص است و در خوشه‌بندی با اتکا به خصوصیات موجود در داده‌ها اقدام به دسته‌بندی می‌کنیم. خوشه‌بندی عضوی از مجموعه‌ی یادگیری بدون ناظر[۴۰] است که شامل تمام مطالعاتی است که در آن‌ها فرض قبلی در مورد برچسب طبقات داده‌ها برقرار نیست. برای اطلاعات کامل در مورد انواع خوشه‌بندی و یادگیری بدون ناظر به کاسامبرا[۴۱] [۲۱] مراجعه شود.
- کاهش بُعد[۴۲]، مبحثی درزمینه‌ی تلخیص داده‌هاست. برخی داده‌های بُعد بالا، ایجاد نقایصی در خواص ریاضیاتی مدل‌ها را در بر خواهند داشت. کاهش بعد یک مبحث مشترک در آمار و جبرخطی است که کاربرد در تلخیص داده‌ها دارد.

- یادگیری ماشین[۴۳]، اساساً ابزار محاسباتی کامپیوتر محور در علم داده‌ها محسوب می‌شود که با ارائه‌ی الگوریتم‌های محاسباتی دقیق و پرسرعت موجب تسهیل دریافت خروجی می‌شود. درواقع، یادگیری ماشین کاربردی از هوش مصنوعی است که به سیستم، توانایی یادگیری خودکار و بهبود تجربه را می‌دهد. جهت اطلاعات بیشتر در مورد فن‌های یادگیری ماشین به بیشاپ [۴۴] [۳] مراجعه شود.

آنچه در بالا مطرح شد ابزارهای مورداستفاده در علم داده‌ها بود. در ادامه از دیدگاه آماری به این ابزارها نگاه خواهیم کرد.

۳. آمار و نقش مبانی نظری آن در علم داده‌ها

بر اساس موارد ذکرشده در بخش ۲، علم داده‌ها از رگرسیون، کلاس‌بندی آماری، خوشه‌بندی، کاهش بعد و یادگیری ماشین برای تحلیل، پیش‌بینی و ارائه خروجی در مورد داده‌ها استفاده می‌کند. همان‌طور که در زیربخش ۳.۲ ذکر شد، تمامی موارد فوق (غیر از الگوریتم‌های رایانه‌ای در یادگیری ماشین) اساساً مباحثی ریشه‌ای در علم آمار هستند. درواقع این‌چنین استنتاج می‌شود که در علم داده‌ها از یادگیری ماشین استفاده می‌کنند تا مباحث نظری آماری را پیاده‌سازی کنند. اگر تاکنون به پاسخ نرسیده‌اید، اکنون سؤال دیگری مطرح می‌کنیم که

---

[37] Clustering

[38] Driver

[39] Kroeber

[40] Unsupervised Learning

[41] Kassambara

[42] Dimension Reduction

[43] Machine Learning

[44] Bishop

دریافتن پاسخ کمک خواهد کرد.

فرض کنید مسافرانی سوار هواپیما شوند، آیا مسافران را هواپیما به مقصد رساننده است یا خلبان؟ بهتر مطرح کنیم: آیا هواپیما ابزاری در دست خلبان برای جابجایی مسافران بوده یا خود آن‌ها را به مقصد رساننده است؟ مسافران همان داده‌ها، خلبان همان نظریات آماری و علم داده‌ها همان هواپیمایی است که آمار، داده‌ها را به کمک آن به مقصد که همان تحلیل و نتایج است می‌رساند.

هوش مصنوعی به‌عنوان یکی از ابزارهای پیشرفته‌ی مورداستفاده‌ی علم داده‌ها شناخته می‌شود. گودفلوو[۴۵] و همکاران [۱۶] در مورد نقش هوش مصنوعی در علم داده‌ها این‌گونه تعبیر می‌کنند که، فن‌های مرتبط با تحلیل و محاسبات در هوش مصنوعی معمولاً بر پایه‌ی یادگیری عمیق[۴۶] طراحی‌شده‌اند و شدیداً تشنه‌ی داده هستند. نقش هوش مصنوعی در علم داده‌ها را می‌توان تشبیه به یک رایانه کرد که وظایف یک کارشناس آمار را انجام می‌دهد. درواقع مباحث نظری آمار به‌وسیله علم داده‌ها روی داده‌های حجیم تولیدشده در عصر حاضر پیاده‌سازی می‌شوند. می‌توان گفت هوش مصنوعی نکته‌ای شبیه به خلبان خودکار در هواپیماست. یعنی از ماهیت موضوع که فاعل اصلی آمار است، چیزی کاسته نمی‌شود. در ادامه، به بررسی بیشتر پاسخ‌های ارائه‌شده در مورد رابطه‌ی بین آمار و علم داده‌ها می‌پردازیم. درنهایت می‌خواهیم به پاسخی تجمیعی برای سؤالی که از ابتدا مطرح کرده‌ایم، برسیم.

۴. آیا علم داده‌ها چیزی بیش از آمار است یا تنها نسخه‌ی پیشرفته‌ی آمار می‌باشد؟

در این بخش، به بررسی و نقد نظرات برخی از افراد صاحب‌نظر در دنیای آمار و علم داده‌ها درباره‌ی ارتباط این دو رشته می‌پردازیم:

- چمبرز [۸] در ۱۹۹۳، در انتقاد به واکنش آماردان‌ها به پیشرفت فناوری این‌گونه بیان داشته که:
آمار در پاسخ به فناوری‌های جدید تغییرات قابل‌توجهی نداشته است. این رشته بر تئوری تأکید دارد و دروس آمار مقدماتی بیشتر تمرکز بر آزمون‌های آماری دارد تا محاسبات آماری.
نقد و بررسی: با توجه به این‌که این نظرات در دهه ۱۹۹۰ بوده است؛ نقدی قابل‌پذیرش و نوعی هشدار در جهت همسو شدن فعالان رشته‌ی آمار با پیشرفت فنّاوری است. به‌صورت غیرمستقیم و البته آنچه در حقیقت رخ داد نیز این بود که در شرایط احساس نیاز به تحلیل داده‌های حجیم، علمی به اسم علم داده‌ها در بستر نظری دانش آمار ظهور کرد.

- کیلیولند [۹] در مورد ظهور علم داده‌ها این‌گونه بیان می‌کند که:
علم داده‌ها را باید بر اساس میزانِ توانمندسازی یک تحلیلگر در تفسیر داده‌ها قضاوت کرد. ابزارهایی

[45] Goodfellow

[46] Deep Learning

که توسط تحلیلگر داده‌ها استفاده می‌شود، اثربخشی مستقیم دارند. نظریه‌هایی که به‌عنوان مبنایی برای توسعه ابزارها عمل می‌کنند، فواید غیرمستقیم دارند.
نقد و بررسی: نظر کیلیولند [۹] تبیین‌کننده‌ی این نکته است که علم داده‌ها در بستر نیاز به گسترش روش‌های تحلیل داده‌ها به وجود آمده است. وی ابزارهای تحلیل را مستقیماً مؤثر بر علم داده‌ها دانسته و مبانی نظری را به‌صورت غیرمستقیم برای این علم سودمند دیده است. این نظر منافاتی با این‌که علم داده‌ها درواقع گسترش روش‌های آماری و مبانی نظری آن در جهت حل مسائل جدید است، ندارد.

- لوکیدز[47] [۲۲] معتقد است که تنها تفاوت میان آمار و علم داده‌ها این است که علم داده‌ها از رویکردهای کلی‌نگرتری استفاده می‌کند.
نقد و بررسی: نظر لوکیدز [۲۲] همسو با این عقیده است که علم داده‌ها از دیدگاه مبانی تئوری از آمار جدا نیست. درواقع وی تنها روش‌های حل مسئله میان این دو رشته را متفاوت می‌داند.

- سیلور[48] آماردان و ژورنالیست مشهور آمریکایی درباره ارتباط علم داده‌ها با آمار این‌گونه بیان می‌کند که علم داده‌ها تنها می‌تواند نام دیگر آمار باشد نه رشته‌ی علمی جداگانه [۲۷].
نقد و بررسی: نظر سیلور [۲۷] بر این اساس است که علم داده‌ها از همان مبانی نظری مورداستفاده‌ی آمار برای تحلیل نتایج استفاده می‌کند. او معتقد است که آنچه از آمار نیاز داریم چیزی جدا از علم داده‌ها نیست.

- برومن[49] [۵] در مقاله‌ای تحت عنوان علم داده‌ها همان آمار است این‌گونه بیان کرده است که وقتی در فیزیک از ریاضی استفاده می‌کنیم فیزیکدانان نمی‌توانند بگویند در حال انجام علم اعداد هستند، آن‌ها دارند از ریاضی استفاده می‌کنند. بنابراین، هرگاه ما تجزیه‌وتحلیل انجام دهیم، در حال بهره بردن آمار هستیم. شما می‌توانید آن را علم داده‌ها یا انفورماتیک یا تحلیل یا هر چیز دیگری بنامید، اما هنوز هم آمار هست و این نباید باعث شود اصطلاح «آمار» را کنار بگذارید.
نقد و بررسی: دیدگاه برومن [۵] همراستا با نظر سیلور [۲۷] است. درواقع هر دو بر این باورند که علم داده‌ها چیزی جدای از آمار نیست و تحلیل داده‌ای که علم داده‌ها بدان می‌پردازد درواقع استفاده از آمار است.

---
[47] Loukides

[48] Silver

[49] Broman

- گلمان[50] [۱۵] بر این باور است که آمار کمترین نقش را دارد و آمار را یک بخش غیراساسی علم داده‌ها برشمارده است.

نقد و بررسی: بر اساس موارد قیدشده در بخش ۲، نقش به سزای مباحث نظری آمار در ساختن پیکره‌ی ابزارها و فن‌های مورداستفاده برای تحلیل داده در علم داده‌ها غیرقابل‌انکار است. درواقع علم داده‌ها بدون رگرسیون، سری زمانی، کلاس‌بندی و خوشه‌بندی چیزی بیشتر از یک ماشین‌حساب پرسرعت نخواهد بود و این فن‌های مذکور تماماً از سرفصل‌های شناخته‌ی شده‌ی علم آمار هستند. نظر نگارنده بر آن است که گلمان [۱۵] از دیدگاه توان محاسباتی و نقطه‌نظر فناوری‌های رایانه‌ای به مسئله پرداخته‌اند. در کل نقطه‌ی قوت علم داده‌ها استفاده از الگوریتم‌های پیچیده و پیشرفته‌ی کامپیوتری است که همین الگوریتم‌ها نیز درواقع کدنویسی از فن‌های آماری روی کاغذ هستند. بدون مباحث نظری آمار روی کاغذ این الگوریتم‌ها وجود خارجی ندارند.

- دهار[51][۱۱] دانشمند هندی، معتقد است که آمار تأکید بر داده‌های کمی و توصیف و تفسیر آن‌ها دارد و علم داده‌ها در تعامل با داده‌های کمی و کیفی سعی و تأکید بر تحلیل و پیش‌بینی این داده‌ها دارد.

نقد و بررسی: نظرات آقای دهار عملاً تعابیری نادرست از دانش آمار مطرح می‌کند. بر اساس هستی و همکاران [۱۹]، پیش‌گویی و پیش‌بینی بر اساس مدل‌بندی داده‌ها به کمک روش‌های آماری چه برای داده‌های کمی و چه کیفی امکان‌پذیر است. مسئله‌ی رگرسیون عملاً مدل‌بندی و پیش‌بینی داده‌هایی از متغیر هدف است که بر اساس ویژگی‌های موجود و تعیین مدل مناسب این هدف در دسترس قرار می‌گیرد. همچنین ارائه‌ی مدلی که داده‌ها را به کلاس‌های گوناگون طبقه‌بندی کند و یک ورودی جدید را به یک طبقه‌ی خاص نسبت دهد عملاً رویه‌ای برای تحلیل داده‌های کیفی است که در مسئله‌ی کلاس‌بندی آماری انجام می‌شود. فن‌های یادگیری ماشین و الگوریتم‌های موجود در آن، این امکان را به روش‌های مذکور می‌دهد که نتایج و تحلیل‌ها را با سرعت بهتر انجام دهند. همچنین امکان تحلیل داده‌های حجیم با استفاده از زبان برنامه‌نویسی پایتون و به کار بستن روش‌های آمار مذکور وجود دارد. شایان‌ذکر است که در دنیای امروز به کار بستن یک رویکرد علمی آماری در یک ماشین یادگیری و گرفتن خروجی نهایی را از وظایف علم داده‌ها می‌دانند. حال‌آنکه بر اساس آنچه گفته شد تنها یک ابزار محاسباتی قوی در خدمت روش‌ها و مفاهیم نظری آمار داریم.

[50] Gelman

[51] Dhar

- دونوهو[۵۲] [۱۳] در یک کارگاه تخصصی در متنی تحت عنوان ۵۰ سال از علم داده‌ها، اذعان داشته است که علم داده‌ها به‌واسطه‌ی بهره جستن از فن‌های محاسباتی پیشرفته و حجم داده‌های مورد تحلیل آن‌که از آمار حجیم‌تر است؛ از آمار متمایز نمی‌شود. او بر این باور است که علم داده‌ها تنها یک زمینه‌ی کاربردی است که در بستر آمار سنتی رشد کرده است.

نقد و بررسی: بر اساس نظر دونوهو [۱۳]، علم داده‌ها پدیده‌ای است که بر اساس دانش آمار طرح‌ریزی‌شده و علم داده‌ها را یک زمینه‌ی کاربردی می‌داند که از دانش آمار نشأت می‌گیرد. این تعبیر دونوهو چیزی دور از این جمله نیست که علم داده‌ها فرزند آمار است و عملاً پدیده‌ای است که آمار را آماده‌ی پاسخ‌دهی به نیازهای متناسب با زمانه یعنی تحلیل داده‌های حجیم و بهره جستن از الگوریتم‌های رایانه‌ای کرده است.

درواقع خالی از ارزش نیست اگر بگوییم، علم داده‌ها تنها نسخه‌ای جدید از آمار است با ابزارهای پیشرفته‌ی محاسباتی که متناسب با نیازهای روز پژوهشگران در علوم کاربردی عمل می‌کند. این نسخه‌ی جدید از آمار را می‌توان آمار+ یا همان Statistics+ در نظر گرفت.

۵. نتیجه‌گیری

بر اساس موارد گفته‌شده، علم داده‌ها و آمار هر دو یک مقصد دارند. پیش‌بینی و تحلیل و تفسیر داده‌ها هدف اصلی هر دو است. داده‌های حجیم و نیاز مبرم به فن‌های محاسباتی پیشرفته برای تحلیل و مدل‌بندی آن‌ها ازجمله اصلی‌ترین انگیزه‌های ظهور گرایشی علمی به نام علم داده‌ها است. علم داده‌ها از همان مفاهیم و مبانی نظری موجود در آمار بهره می‌برد. بزرگ‌ترین وجه تمایز آمار و علم داده‌ها ابزارهای نوین نرم‌افزاری در تحلیل و مدل‌بندی داده‌ها است. علم داده‌ها با اتکا به مبانی آمار و ابزار نرم‌افزاری مدرن به تحلیل داده‌هایی می‌پردازد که در قرن ۲۱ به‌واسطه پیشرفت صنعت و فنّاوری تولیدشده‌اند. اگر این‌گونه مطرح کنیم که علم داده‌ها همان آمار است که مبانی نظری آن را در نرم‌افزارهای پیشرفته پیاده‌سازی شده است؛ درواقع می‌توان نتیجه گرفت که علم داده‌ها درواقع نسخه‌ی پیشرفته‌ی علم آمار یا همان آمار+ است.

۶. تشکر و قدردانی

نویسندگان این مقاله، از داوران محترم بابت ارتقاء سطح کیفی مقاله با نظرات سازنده و مفیدشان و همچنین از دبیر رابط و سردبیر محترم نشریه بابت مد نظر قرار دادن مقاله کمال قدردانی و تشکر خود را ابراز می‌دارند.

مراجع

[1] A. Agresti, *Categorical data analysis,* (Vol. 792), John Wiley & Sons, 2012.

[52] Donoho

[2] A. Bhardwaj, *What's the Difference Between Data Science and Statistics?* Priceonomics. https://priceonomics.com/whats-the-difference-between-data-science-and/(2015).

[3] C.M. Bishop, *Pattern recognition and machine learning,* Springer, New York, 2006.

[4] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and regression trees,* Chapman and Hall/CRC, 1984.

[5] K. Broman, *Data science is statistics,* The Stupidest Thing: Statistics, Genetics, Programming, Academics, 2013.

[6] I. Carmichael and J.S. Marron, *Data science vs. statistics: two cultures?* Jpn. J. Stat. Data Sci, **1**, (2018) 117–138.

[7] G. Casella, S. Fienberg and I. Olkin, *Springer Texts in Statistics,* Design (Vol. 102) 2006.

[8] J.M. Chambers, *Greater or lesser statistics: a choice for future research,* Statist. Comput., **3** (1993), no. 4, 182–184.

[9] W.S. Cleveland, *Data science: an action plan for expanding the technical areas of the field of statistics*, International statistical review, **69** (2001) no. 1, 21–26.

[10] R.D. De Veaux, M. Agarwal, M. Averett, B.S. Baumer, A. Bray, T.C. Bressoud, L. Bryant, L.Z. Cheng, A. Francis, R. Gould and A.Y. Kim, *Curriculum guidelines for undergraduate programs in data science.* Annu. Rev. Stat. Appl., **4** (2017) 15–30.

[11] V. Dhar, *Data science and prediction, Communications of the ACM*, **56** (2013) no. 12, 64–73.

[12] P.J. Diggle, *Statistics: a data science for the 21st century,* J. R. Stat. Soc., A: Stat. Soc. **178** (2015) no. 4, 793–813.

[13] D. Donoho, *50 Years of Data Science*, J. Comput. Graph. Stat. **26** (2017) no. 4, 745–766.

[14] H.E. Driver and A.L. Kroeber, *Quantitative expression of cultural relationships* Berkeley, University of California Press, 1932.

[15] A. Gelman, *Statistics is the least important part of data science,* Statistical Modeling, Causal Inference, and Data Science blog. 2013.

[16] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning* (Vol. 1) USA: MIT Press, 2017.

[17] A. Goodman, C. Kamath and V. Kumar, *Data analysis in the 21st century,* Statistical Analysis and Data Mining: The ASA Data Science Journal, **1** (2008) no. 1, 1–3.

[18] J. Hardin, R. Hoerl, N.J. Horton, D. Nolan, B. Baumer, O. Hall-Holt, P. Murrell, R. Peng, P. Roback, D. Temple Lang and M. Ward, *Data science in statistics curricula: Preparing students to "think with data",* The American Statistician **69**(2015) no. 4, 343–353.

[19] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning (2nd Edition),* New York: Springer, 2009.

[20] S.C. Hicks and R.A. Irizarry, *A guide to teaching data science,* The American Statistician, **72**(2018) no.4, 382–391.

[21] A. Kassambara, *Practical guide to cluster analysis in R: Unsupervised machine learning,* France: Statistical Tools For High-Throughput Data Analysis, 2017.

[22] M. Loukides, *What is data science?* O'Reilly Media, Inc. 2011.

[23] D. Nolan and D. Temple Lang, *Computing in the statistics curricula,* The American Statistician, **64**(2010) no. 2, 97–107.

[24] P. Pham, *The Impacts of big data that you may not have heard of,* Forbes. https://www.forbes.com/sites/peterpham/2015/08/28/the-impacts-of-big-data-that-you-may-not-have-heard-of/(2015).

[25] G. Press, *A very short history of big data,* Forbes Tech Magazine. https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/(2013).

[26] N. Reid, *Statistical science in the world of big data,* Stat. Probab. Lett. **136**, (2018) 42–45.

[27] N. Silver, *What I need from statisticians,* Stats & Data Science Views. https://www.statisticsviews.com/article/nate-silver-what-i-need-from-statisticians/(2013).

[28] D. Singpurwalla, *A handbook of statistics: An overview of statistical methods*, 2013.

[29] D. Von Winterfeldt, and W. Edwards, *Decision analysis and behavioral research,* Cambridge University Press, Cambridge, 1986.

[30] C. Weihs and K. Ickstadt, *Data science: the impact of statistics,* Int. J. Data Sci. Anal. **6** (2018) 189–194.